



A Weighted Hybrid of TF-IDF and Wining for Content-Based Scientific Article Recommendation with Citation-Based Ground Truth

Moch Zawaruddin Abdullah^{1*}, Krisna Andika Wijaya², Imam Fahrur Rozi³, Endah Septa Sintiya⁴,
Dian Septiani Santoso⁵

^{1,2,3,4}Department of Information Technology, Politeknik Negeri Malang, Indonesia

⁵Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia

Abstract.

Purpose: Locating the few papers that bear on a research question gets harder as an institutional repository grows. The recommender built here scores each candidate twice over: TF-IDF with Cosine Similarity weighs terms against the whole corpus, while Wining with Jaccard Similarity compares local character fingerprints.

Methods: Titles and abstracts of 1,954 host-institution articles passed through case folding, filtering, tokenizing, stop-word removal, and Sastrawi stemming. A single joint grid of 462 combinations of K-gram length, window size, and fusion weight was scored on Precision, Recall, F1-Score, MAP, and NDCG at cut-offs 5, 10, and 20 against a citation-based ground truth of 31 evaluable queries; the winner was then checked by nested leave-one-query-out cross-validation, against a BM25 baseline, and with paired bootstrap confidence intervals.

Result: K = 9, W = 6 and a 60:40 TF-IDF to Wining split led under every selection criterion and in all 31 folds, at MAP@5 = 0.3567, MAP@20 = 0.3745, NDCG@20 = 0.4613. The lead over TF-IDF alone is 0.018 MAP@20 (95% CI 0.004 to 0.035), a small margin. BM25 reached 0.3233, Wining alone 0.2316.

Novelty: One lexical-statistical channel and one structural channel are fused by a weight read off an exhaustive grid instead of set by hand, and the whole thing is tested on a local institutional corpus whose relevance labels come from citation links that a reader can rebuild from public metadata.

Keywords: Recommendation system, TF-IDF, Cosine similarity, Wining, Jaccard similarity

Received July 2026 / **Revised** September 2026 / **Accepted** October 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



INTRODUCTION

Digital publication output keeps growing, and the way researchers reach the literature has changed with it. Repositories now hold far more than any reader can scan, so locating the handful of references that match a research focus takes longer than it used to [1], [2]. The cost is twofold: time spent searching instead of writing, and sources missed because a keyword did not line up. Conventional search engines compound the problem, since they match surface terms and lose much of the context an article carries [3].

Recommender systems filter that flood. They point a researcher toward relevant papers, datasets, and possible collaborators, which returns hours otherwise lost to searching [2], [4]. Scientific article recommendation is now an active research area, and hybrid designs that read more than one content-based signal are the ones reporting the strongest results [5], [6].

Certain efforts have been made to study content-based matching in academic settings. Arsad et al. [7] employed text mining with Cosine Similarity to estimate document similarity of undergraduate thesis documents which was proved to be efficient for high-sparsity text corpora. Amin et al. [8] established a Journal Finder based on Jaccard Similarity for the titles and abstracts of articles, however basic metadata matching does not allow for adaptation to different text forms. More recently, hybrid designs, such as those incorporating semantic and lexical signals [6], [9]–[11], have been shown to perform better than single-method baselines in Precision@K and NDCG@K.

^{1*}Corresponding author.

Email addresses: zawaruddin@polinema.ac.id (Abdullah).

DOI: [10.15294/sji.v13i4.59181](https://doi.org/10.15294/sji.v13i4.59181)

Winnowing method paired with Jaccard Coefficient has been found to be robust for lexical similarity text-matching tasks within the subdomain. Zachrias and Gunawan [12] showed that the combination may achieve good precision and recall in detecting plagiarism, but Bramantya et al. [13] noted that the detection sensitivity is highly dependent on the choice of K-gram and window size values. Despite this progress, research that integrates the Winnowing-Jaccard pair with the TF-IDF-Cosine pipeline into a unified weighted scoring mechanism for recommendation remains limited, particularly on faculty-scale institutional corpora. The gap is threefold. First, each of the closest studies relies on one channel only. Arsad et al. [7] and Amin et al. [8] use a single similarity measure, on thesis documents and on journal metadata respectively, while Zachrias and Gunawan [12] and Bramantya et al. [13] use Winnowing-Jaccard for plagiarism detection, not for recommendation; none of these works tests the two pairs together in one ranking function. Second, the hybrid content-based recommenders that do combine several signals [6], [9]–[11] fuse embedding-derived channels and set their fusion weights heuristically, leaving the weight itself unvalidated. Third, none of these studies evaluates on an institutional corpus against relevance judgments derived from citation links, so their results cannot be reproduced by a third party without access to the original user data. This study addresses all three points: it pairs a lexical-statistical channel with a structural-fingerprint channel, derives the fusion weight from an exhaustive grid search instead of by hand, and evaluates the result on a citation-based ground truth that any reader can reconstruct from public metadata.

Concretely, the recommender scores each candidate with one weighted linear interpolation of TF-IDF-Cosine Similarity and Winnowing-Jaccard Similarity. The paper makes three contributions: (i) a dual-channel hybrid recommender built and validated on a local institutional corpus that had not been studied before, namely 1,954 articles from the host institution [5]; (ii) fusion weights found by a systematic grid search over K-gram, window size, and weight, instead of weights set ad hoc [14]; and (iii) an evaluation whose relevance labels come from citations, an objective proxy for academic relevance that others can reproduce [15]–[17].

Corpus quality bounds what any recommender can return. Lghaouch et al. [18] filtered internally inconsistent documents out of a collection and saw ranking improve with the recommender backbone untouched. This study takes the same corpus-first view, pairing a lexical similarity channel with a document-fingerprint channel over an institutional publication set.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing procedures, scoring methods, and evaluation protocol. Section 3 presents the experimental results, including analyses of K-gram and window size, hybrid weighting, citation-based evaluation, and testing with student proposals. Section 4 concludes the paper and outlines directions for future work.

METHODS

Figure 1 outlines the whole process. Articles are collected and de-duplicated, go through one shared pre-processing stage, and are then scored by the two similarity channels and their fusion; the two evaluation tracks are explained later in this section.

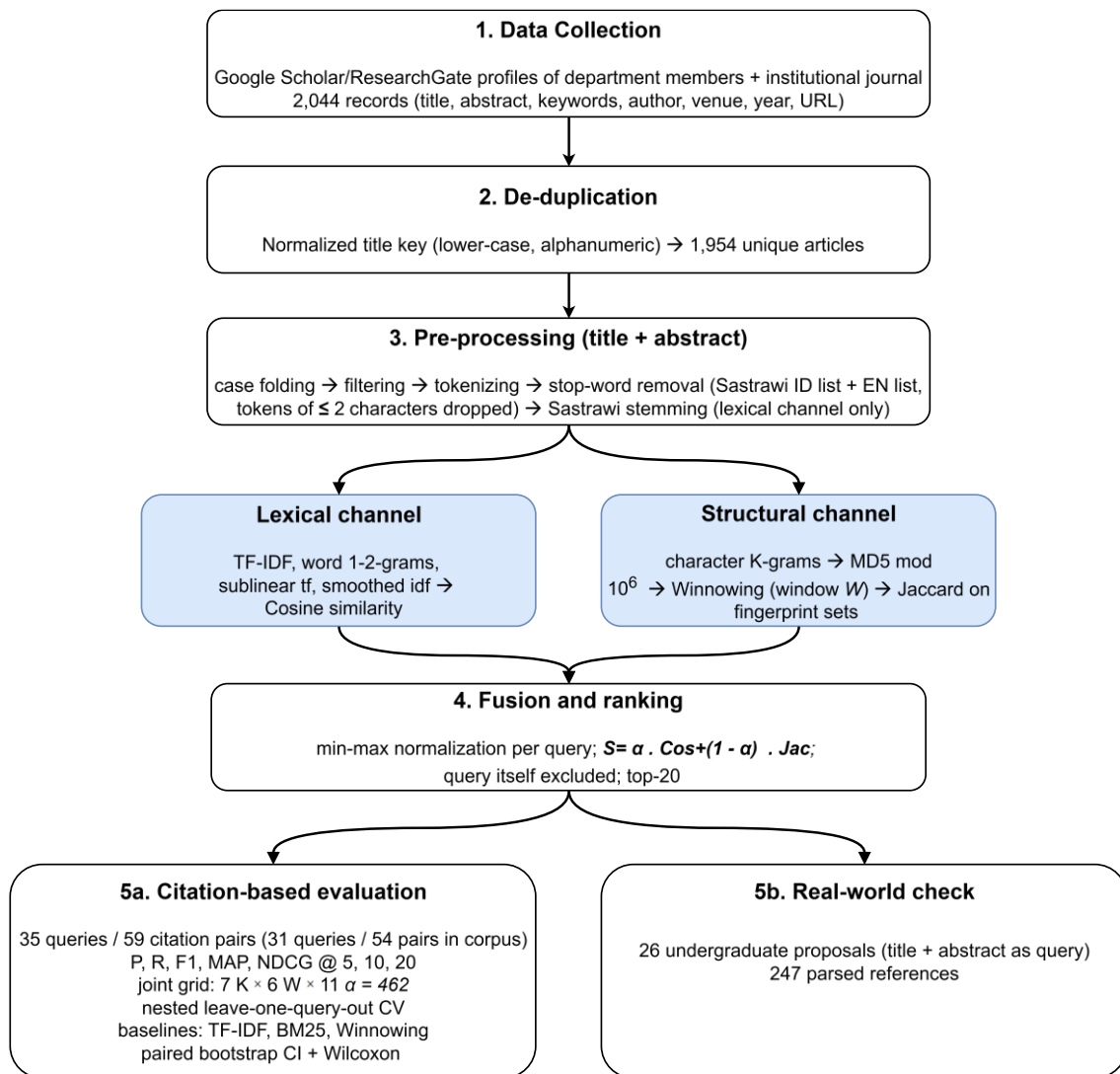


Figure 1. Research process of the dual-channel recommender and its evaluation

Dataset

The articles were harvested from open-access sources, mainly ResearchGate and Google Scholar, through the publication profiles of the faculty members of the host institution. We collected 1,702 records from these profiles and 342 articles from the host institution's own journal (2020 to 2025), 2,044 records altogether. After duplicates were removed on a normalized title key (lower-cased, all non-alphanumeric characters stripped), 1,954 unique articles remained. Their topics include content-based paper recommendation [1], web scraping and harvesting workflows [19], hybrid recommender-system evaluation [6], and many other subjects in Information Systems, Machine Learning, and Informatics. Each record holds the title, abstract, keywords, authors, journal name, publication date, source URL, and the DOI where one exists. The citation-based evaluation reads only the title and abstract. For the 31 articles that have no abstract, the title carries the query on its own.

The profile-based workflow proposed by Rahmatulloh and Gunawan [20] was adopted in this study. Researcher profiles and publication attributes were extracted from Google Scholar through the HTML DOM. The same approach was subsequently reused in two studies: structured article lists were constructed by Fahrudin et al. [21], while the approach was integrated into a recommender system based on user history by Utama et al. [22].

Text Preprocessing

Query articles and corpus articles meet the scoring stage in the same shape, because both pass the five stages below:

1. *Case Folding*: Text is lower-cased, so a term written with different capitalization is not counted twice [23].
2. *Filtering*: The lexical channel replaces every non-alphabetic character with a space; the structural channel keeps digits, and once stop words are gone it strips whitespace and punctuation as well, which leaves the unbroken character stream the fingerprints run over [24].
3. *Tokenizing*: *Whitespace marks the token boundaries* [23].
4. *Stop-word Removal*: Out go the tokens on the Sastrawi Indonesian stop-word list, which we extended with 45 English function words and generic domain terms such as data, system, and method; tokens of two characters or fewer go with them [19].
5. *Stemming*: Inflected Indonesian words are reduced to their base forms using the Sastrawi library, which implements the Nazief–Adriani algorithm with confix-stripping rules; stemming has been shown to raise the accuracy of downstream classification on Indonesian text [25]. Stemming is applied to the lexical channel only; the structural channel fingerprints the unstemmed surface form, which is what makes it sensitive to shared phrasing rather than shared vocabulary.

For Indonesian-language corpora, Maylawati et al. [26] emphasize the importance of preprocessing flexibility at the library level to systematically handle morphological variants, while Suzanti and Jauhari [27] showed that the choice of stemmer (Porter, Nazief–Adriani, or Enhanced Confix Stripping) changes retrieval results under both Cosine and Dice similarity.

TF-IDF and Cosine Similarity

In the lexical channel, document similarity is computed in a vector space model over word unigrams and bigrams (170,951 features). Term Frequency is scaled sublinearly, so that a term repeated many times does not dominate the vector, while Inverse Document Frequency downweights terms that are common across the corpus; each document vector is then L2-normalized [28], [29].

$$TF(t, d) = 1 + \ln f(t, d) \quad (1)$$

$$IDF(t) = \log\left(\frac{N+1}{df(t)+1}\right) + 1 \quad (2)$$

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t) \quad (3)$$

where $f(t, d)$ is the raw count of term t in document d , N is the number of documents in the corpus, and $df(t)$ is the number of documents containing t ; the added ones keep the ratio defined for every term. Cosine Similarity then measures the proximity between two TF-IDF vectors [29], [30]:

$$\text{Cosine}(A, B) = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}} \quad (4)$$

where A and B are the TF-IDF vectors of the query document and the corpus document. Because all weights are non-negative, values lie between 0 (no shared term) and 1 (identical vectors) [28].

Winnowing and Jaccard Similarity

The structural channel applies the Winnowing algorithm [31], which generates document fingerprints in three stages: K-gram generation, hashing, and minimum-hash selection within a sliding window. Each pre-processed text is turned into the sequence of its overlapping character substrings of length K ; each K-gram is hashed with MD5 and reduced modulo 10^6 ; and within each window of W consecutive hashes the minimum is kept. The fingerprint of a document is the set of these minima, without positional information [12], [13].

Parameter selection plays a critical role in the winnowing process. Schleimer et al. [31] demonstrated that any shared substring with a length of at least $t = W + K - 1$ is guaranteed to produce a common fingerprint. With $K = 9$ and $W = 6$, the resulting threshold is 14 characters. Nasien and Yansen [32] applied winnowing to the titles and abstracts of student final projects, which are comparable in form to the texts used in the

present corpus. In addition, Bramantya et al. [13] showed that the selection of K and window size substantially influences the sensitivity of winnowing for Indonesian text. After fingerprint extraction, the similarity between two fingerprint sets was measured using the Jaccard Coefficient [33], [34]:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (5)$$

The Jaccard Coefficient ranges from 0 to 1, where values closer to 1 indicate a greater degree of overlap between the fingerprint sets of the two documents [7], [33].

Hybrid Score

The scores obtained from the two channels were combined into a final similarity score using weighted linear interpolation [3], [35], [36]:

$$S(q, d_i) = w_1 \cdot \text{Sim}_{\text{Cosine}}(q, d_i) + w_2 \cdot \text{Sim}_{\text{Jaccard}}(q, d_i) \quad (6)$$

subject to $w_1 + w_2 = 1$. Both score vectors of a query are min-max normalized to $[0, 1]$ over the corpus before fusion, and the query article never appears in its own ranking. Two warnings from the literature shaped this step: Rizzo et al. [3] note that a convex combination of two similarity scores looks simple but fails without careful normalization and weight calibration, since scores drawn from unlike distributions cancel one another, and Haubner and Setzer [35] report that the best weight pair moves with the model and the dataset, so we searched the weights rather than fixing them.

We varied w_1 and w_2 in steps of 0.1, from 1.0 : 0.0 down to 0.0 : 1.0, jointly with $K \in \{3, 5, 7, 9, 11, 13, 15\}$ and $W \in \{2, 3, 4, 5, 6, 7\}$, so that $7 \times 6 \times 11 = 462$ configurations were scored in all. Searching the whole grid removes the ad hoc weight choice that Çano and Morisio [6] identify as a recurring weakness in hybrid recommender design.

Evaluation Methodology and Ground Truth Construction

What follows sets out the retrieval metrics, then the citation-based ground truth, then the fixed output size that keeps every run comparable with the baselines.

Evaluation Metrics

Five metrics in standard use for content-based recommenders carry the evaluation: Precision@K, Recall@K, F1-Score@K, MAP@K, and NDCG@K [4], [14], [37], chosen so the numbers here can sit beside published hybrid and single-method baselines [5], [6], [9], [38]. Precision@K is the share of relevant documents inside the top-K list; Recall@K is the share of all relevant documents the cut-off recovers [37]; F1-Score@K is their harmonic mean. MAP@K averages precision over the ranks where relevant documents land, and NDCG@K discounts by rank position, so a relevant document counts for more the higher it sits [4], [14].

Relevance is binary throughout. For a query q , AP@K sums the precision at every relevant rank inside the top K and divides by $\min(|Rq|, K)$, where Rq is the relevant set of q ; MAP@K averages AP@K over queries; NDCG@K uses unit gains with a $\log_2$ rank discount, normalized against an ideal ranking of $\min(|Rq|, K)$ relevant items.

Ground Truth Construction

Following Klemiński et al. [15] and Boyack and Klavans [16], relevance was defined using explicit citation links rather than user ratings, and only direct citations were considered. The reference list of each query article was parsed, and every cited title that matched an article in the corpus after normalization (lower-casing and removing non-alphanumeric characters) was added to the corresponding relevant set. Co-citation and bibliographic coupling, which have also been examined as relatedness signals Sjögarde and Ahlgren [17], were not included in the present evaluation.

This procedure initially produced 59 citation pairs across 35 query articles, with a mean of 1.69 and a median of 1 relevant article per query. However, four of the 35 citing articles were not present in the de-duplicated corpus. The final evaluation therefore included 31 queries and 54 citation pairs, with a mean of 1.74, a median of 1, and a maximum of 11 relevant articles per query. Each query was represented by the corresponding citing article in the corpus, which was excluded from its own ranking. The resulting relevant

sets were sparse and uneven, making the evaluation more restrictive than benchmarks constructed from larger synthetic ground-truth pools [16].

Output Format

For each query, the system returned the 20 articles with the highest hybrid scores, $S(q, d_i)$, as defined in Equation (6). The output size was fixed across all experimental runs to ensure that any observed differences in performance were attributable solely to the evaluated configuration.

Model Selection and Statistical Testing

The configuration is selected by MAP@20, the rank-aware metric at the output cut-off; Table 2 also reports an aggregate score, defined as the mean of the five metrics at the three cut-offs (15 values). Because selection and evaluation share the same 31 queries, the selection was repeated inside a nested leave-one-query-out cross-validation [39]: for each query, the configuration with the highest MAP@20 on the remaining 30 queries was chosen and then scored on the held-out query alone. Three reference systems were run on the same pre-processed tokens: each channel of the hybrid on its own, and Okapi BM25 ($k_1 = 1.5$, $b = 0.75$) over word unigrams as the standard sparse lexical baseline [40]. Differences between systems are reported as the mean per-query difference with a 95% confidence interval from 10,000 paired bootstrap resamples of the queries, together with a two-sided Wilcoxon signed-rank test on the non-zero per-query differences [14].

Implementation Note

The system is implemented in Python with Laravel for the web interface; the text-computation module is decoupled through a command-line interface invoked via inter-process communication. This architectural separation enables independent deployment of the web layer and the recommendation model, but is not part of the scientific contribution.

RESULT AND DISCUSSION

K-Gram Value Testing

In the first experiment, the K-gram length and the window size were searched jointly with all eleven fusion weights: $K \in \{3, 5, 7, 9, 11, 13, 15\}$ and $W \in \{2, \dots, 7\}$, 462 configurations in total. Table 1 reports, for each K, the window size and weight that maximize MAP@20, and Figure 2 shows the full $K \times W$ surface.

Table 1. Best configuration for each K-Gram value (W and weight selected by MAP@20)

K-gram	Optimal Weight (TF:Win)	Window Size (W)	MAP@5	MAP@10	MAP@20
3	0.9 : 0.1	4	0.3435	0.3526	0.3658
5	0.8 : 0.2	7	0.3339	0.3477	0.3617
7	0.9 : 0.1	5	0.3355	0.3468	0.3624
9	0.6 : 0.4	6	0.3567	0.3625	0.3745
11	0.8 : 0.2	5	0.3409	0.3468	0.3618
13	0.7 : 0.3	5	0.3435	0.3497	0.3638
15	0.8 : 0.2	5	0.3435	0.3494	0.3645

$K = 9$ with $W = 6$ gives the best result at every cut-off (MAP@5 = 0.3567, MAP@10 = 0.3625, MAP@20 = 0.3745). With K between 3 and 7, the best configurations put 80–90% of the weight on TF-IDF and gain no more than 0.009 MAP@20 over TF-IDF alone (0.3567). Short character segments occur in many unrelated documents, and their fingerprints end up marking accidental overlaps instead of shared phrasing [13]. At $K = 9$ Winnowing receives 40% of the weight. Figure 2 shows that this is not a single lucky cell; with $K = 9$, the window sizes $W = 5, 6$, and 7 all reach MAP@20 between 0.3690 and 0.3745. For $K = 11$ to 15 the optimum moves back towards TF-IDF (70–80%), as longer segments become too specific to survive even light paraphrasing, in line with the observation of Schleimer et al. [31] that detection depends jointly on K and window size. Many cells of Figure 2 equal 0.3567 exactly; there, no share of Winnowing improves on TF-IDF alone.

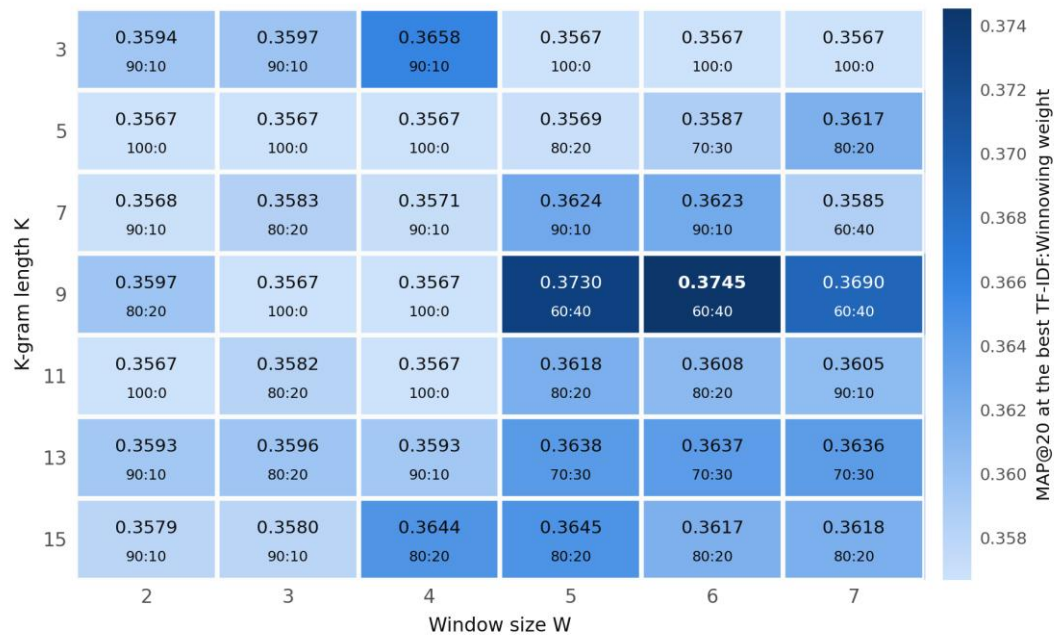


Figure 2. MAP@20 over the joint $K \times W$ grid. Each cell shows the best TF-IDF : WInnowing weight for that pair; the outlined cell is the selected configuration ($K = 9$, $W = 6$, 60 : 40)

Hybrid Weight Testing

With K-gram = 9 and $W = 6$ fixed as the best configuration, the second experiment varied the weight between TF-IDF and WInnowing. Table 2 reports the aggregate score, calculated as the mean of the five evaluation metrics (Precision@K, Recall@K, F1-Score@K, MAP@K, and NDCG@K) at cut-offs 5, 10, and 20 for each weight combination. Although Tables 1 and 2 present the search in two steps, the search itself was joint: over all 462 configurations, $K = 9$, $W = 6$ with a 60:40 weight ranks first under MAP@20, MAP@5, NDCG@10, the mean of F1, MAP, and NDCG, and the aggregate score alike, so the choice of criterion does not change the selected configuration.

Table 2. Aggregate score of each weight combination ($K = 9$, $W = 6$)

Weight (TF-IDF : WInnowing)	Aggregate Score
100% : 0%	0.2989
90% : 10%	0.3038
80% : 20%	0.3100
70% : 30%	0.3099
60% : 40%	0.3180
50% : 50%	0.3071
40% : 60%	0.3080
30% : 70%	0.3028
20% : 80%	0.2919
10% : 90%	0.2711
0% : 100%	0.2467

At 60% TF-IDF and 40% WInnowing the aggregate score reaches 0.3180, against 0.2989 for TF-IDF alone and 0.2467 for WInnowing alone.

N-Gain Analysis

Normalized Gain puts the hybrid against the single-method baseline on one scale:

$$G = \frac{\text{Score}_{\text{hybrid}} - \text{Score}_{\text{baseline}}}{1.0 - \text{Score}_{\text{baseline}}} \quad (7)$$

Table 3. Normalized gain of the hybrid configuration (K=9, W=6)

Metric	Baseline (TF-IDF 100%)	Hybrid (TF : Win = 0.6 : 0.4)	N-gain
MAP@5	0.3315	0.3567	0.038
MAP@10	0.3428	0.3625	0.030
MAP@20	0.3567	0.3745	0.028

N-gain is positive at every cut-off: 0.038 at MAP@5, 0.030 at MAP@10, 0.028 at MAP@20. The structural channel earns its keep at the head of the list, where a misplaced document costs a reader most, and loses ground as the list lengthens [3], [35]. In absolute terms these are small numbers. Whether they survive resampling of the query set is what the robustness checks below decide.

Two earlier reports point the same way. Alaoui et al. [9] compared filtering methods head to head; Çano and Morisio [6] surveyed the hybrid literature; both put hybrid content-based designs ahead of single-method filters on Precision@K and NDCG@K. Rizzo et al. [3] examined ranking models close to the convex combination used here and caution that even a plain linear blend needs careful normalization once the two score distributions diverge, which is why each channel is min-max normalized per query before fusion and the weight is read off the full grid.

Comparative Context

Kuş et al. [10] offer a point of reference, not a comparison. They fused Word2Vec and LDA over 41,000 arXiv articles and reported Precision@50 = 0.206, Recall@50 = 0.791, and F1-Score@50 = 0.498 for their best configuration. Setting MAP@5 = 0.3567 beside Precision@50 = 0.206 would compare different metrics at different cut-offs, on corpora an order of magnitude apart (1,954 institutional articles against 41,000 arXiv articles), under different definitions of relevance; a defensible head-to-head test would need the same corpus, the same cut-off, the same metric, and the same relevance judgments, and is left to future work [10], [11].

Putra et al. [11] show the same pattern at larger scale with Sentence Transformer retrieval: embedding-based methods lift recommendation quality, at a computational cost this faculty-scale corpus does not justify. The hybrid used here runs without GPU hardware.

Ground Truth Evaluation

Before conducting the performance evaluation, an analysis of the characteristics of the ground truth used as the benchmark for document relevance assessment was carried out. The ground truth was constructed based on citation relationships among scientific articles in the research corpus. Table 4 presents the ground truth statistics obtained.

Table 4. Results of ground truth analysis

Statistic	Value
Total valid citation pairs	59
Total query articles	35
Query articles present in the corpus (evaluated)	31
Citation pairs used in evaluation	54
Average ground truth per query	1.69
Median ground truth per query	1
Minimum ground truth	1
Maximum ground truth	11

The relevant sets are sparse and uneven. A median of 1 puts most queries at one truly relevant document in the corpus, or none [15]–[17]. Citation networks are documented as sparse in this way [17], and the sparsity explains a gap that would otherwise look odd: absolute Precision sits at P@5 = 0.1355 while the rank-aware MAP@5 reaches 0.3567, because a query with one or two relevant documents scores well on MAP as soon as the system puts them near the top, and poorly on Precision whatever it does.

Sparsity also bounds how far the benchmark generalizes. A citation pair counts only when the citing article and the cited article both sit inside the corpus, so the coverage ceiling measured in the Student Proposal Testing section shapes the benchmark as much as it shapes practical use. The 31 evaluable queries are whatever part of the corpus happens to cite itself. That is not a random sample of what a user would search

for, and the figures here should be read as evidence about ranking behavior on internally citing subsets of an institutional corpus, not as an estimate of field-wide retrieval quality.

Robustness Checks

The selected configuration and its advantage were tested three ways on the same 31 queries. Nested leave-one-query-out cross-validation returned $K = 9$, $W = 6$, 60:40 in every one of the 31 folds, which makes the held-out estimate equal to the in-sample one at $\text{MAP@20} = 0.3745$; that agreement lowers the risk of an artefact of tuning inside this benchmark, though generalizability still waits on an independent external test set. Table 5 sets the hybrid against each of its channels and against BM25 over the same queries.

Table 5. Hybrid against its single channels and BM25 ($K = 9$, $W = 6$; mean over 31 queries)

System	P@5	MAP@5	MAP@20	NDCG@20	95% CI of MAP@20
BM25 (Okapi)	0.1032	0.2952	0.3233	0.4062	0.192–0.464
Winnowing-Jaccard only	0.1290	0.2108	0.2316	0.3378	0.138–0.343
TF-IDF-Cosine only	0.1226	0.3315	0.3567	0.4417	0.228–0.497
Hybrid 60:40 (in-sample)	0.1355	0.3567	0.3745	0.4613	0.244–0.508
Hybrid, nested LOQO CV	0.1355	0.3567	0.3745	0.4613	0.244–0.508

TF-IDF-Cosine, not BM25, is the stronger lexical baseline here, at $\text{MAP@20} = 0.3567$ against 0.3233. Adding the Winnowing channel to TF-IDF lifts MAP@20 by 0.018 (95% CI 0.004 to 0.035; Wilcoxon $p = 0.036$; 9 queries improved, 19 unchanged, 3 worse), MAP@5 by 0.025 (95% CI 0.004 to 0.052; $p = 0.094$), and NDCG@20 by 0.020 (95% CI 0.003 to 0.041; $p = 0.051$). Every bootstrap interval excludes zero, yet only MAP@20 clears the paired rank test at the 5% level; the gain is small and consistent, and we claim no more than that.

The lead over BM25 is larger at 0.051 MAP@20 , and less certain: the 95% CI runs from -0.027 to 0.137 and takes in zero. Pairing Winnowing with BM25 in place of TF-IDF returned nothing at all, the best BM25 weighting at 80:20 reaching $\text{MAP@20} = 0.3251$, within noise of BM25 alone. The score distributions explain why. After min-max normalization BM25 holds high scores much deeper into the list, a median mean normalized score of 0.21 across queries over ranks 2 to 21 against 0.07 for TF-IDF, and a fingerprint channel has little room to reorder candidates that are already bunched - the distribution effect Rizzo et al. [3] describe for linear fusion.

Against the best configuration at any other K the margin is thinner again. $K = 3$, $W = 4$, 90:10 reaches $\text{MAP@20} = 0.3658$, leaving the selected setting ahead by 0.009 with a 95% CI of -0.012 to 0.033 and $p = 0.39$. $K = 9$ is the best observed value, not a statistically separated optimum. The paired tests support the structural channel; they do not single out this K .

Student Proposal Testing

Real-world validation was conducted using 26 undergraduate thesis proposals authored by students of the bachelor's program. For each proposal, the title and abstract were submitted as a query to the deployed service, which uses a separately constructed production index of 1,955 articles. The production index also stores keywords and follows the preprocessing, filtering, and title-deduplication rules of the deployed pipeline, while the citation-based experiment works on the de-duplicated 1,954-article snapshot from the Methods section and uses titles and abstracts only; the difference of one document comes from these index-construction rules. With the optimal configuration ($K = 9$, $W = 6$, 60% TF-IDF and 40% Winnowing) the service returned the top 20 recommendations for each proposal, and we compared them with the proposal's reference list. This test is an operational check and is not meant to be compared directly with the citation-based benchmark.

The 26 proposals cite 247 references in total, but only 7 of the cited articles (2.8%) are in the corpus. The rest are books, conference papers, or journal articles that the corpus does not cover. Of those 7, the system placed 6 in the top 20 for the proposal that cited them, a recall of 85.7% on this small subset. With a denominator of seven ($n = 7$) the figure is only indicative; its interval is too wide to compare configurations, and we do not treat it as a performance estimate. The one missed reference belongs to a proposal whose only matching corpus article comes from a quite different subdomain and shares very few tokens with it. A

token-matching recommender is not expected to find such a cross-domain document, as earlier work on queries and corpora from separate subfields also reports [1], [7].

Two things follow from this test. When the corpus did hold a cited document, the system usually found it (6 of 7); the sample is far too small to generalize from, although it points the same way as the citation-based results. The bigger problem is coverage: the faculty's own publications account for only 2.8% of what the students cite, and a better algorithm cannot fix that. The quality of the recommendations depends on how complete and how broad the corpus is, as much as on the method [18].

CONCLUSION

This study built a content-based recommender that fuses a TF-IDF-Cosine channel and a Wining-Jaccard channel in one weighted linear interpolation, and tested it on 1,954 articles from the Information Technology Department of the host institution. A joint grid over 462 combinations of K-gram length, window size, and fusion weight, scored against a citation-based ground truth of 31 evaluable queries, selected $K = 9$, $W = 6$ and a 60:40 split at $\text{MAP}@5 = 0.3567$, $\text{MAP}@20 = 0.3745$, and $\text{NDCG}@20 = 0.4613$; every selection criterion pointed to the same setting, and so did all 31 folds of the nested leave-one-query-out cross-validation, which lowers the risk that the parameters were fitted to the benchmark. The Wining contribution is modest. It lifts $\text{MAP}@20$ by 0.018 over TF-IDF alone (95% CI 0.004 to 0.035) and adds nothing measurable on top of BM25, so the value of fingerprint fusion appears to turn on how the lexical scores are distributed. The department gains a retrieval service over its own publications that runs without embedding models or GPUs, together with a citation-based benchmark on which later recommenders for this corpus can be compared without manual relevance judgments. The limits are plain: (i) 31 evaluable queries leave the confidence intervals wide, (ii) the edge of $K = 9$ over the next-best K sits inside sampling noise, and (iii) only 2.8% of the 247 references in the 26 student proposals exist in the corpus, which caps what the proposal test can say. Work on corpus coverage and on a larger citation benchmark [16], [17] should come before further work on the scoring function, after which semantic representations such as FastText, BERT, or Sentence Transformers [10], [11], [41], [42], a corpus-quality filter in the manner of Lghaouch et al. [18], and richer Indonesian preprocessing [26], [27] become worth testing.

ACKNOWLEDGMENT

During the revision of this manuscript the authors used AI assistants (Claude, Anthropic; ChatGPT, OpenAI) to edit the English text, to check the evaluation scripts, and to re-run the statistical analyses reported in the Results. Every AI-assisted passage, analysis, and reference was reviewed and verified by the authors, who take full responsibility for the content of this article.

REFERENCES

- [1] H. Tang, B. Liu, and J. Qian, "Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation," *Concurr. Comput. Pract. Exp.*, vol. 33, no. 13, 2021.
- [2] Z. Zhang et al., "Scholarly recommendation systems: a literature survey," *Knowl. Inf. Syst.*, vol. 65, no. 11, pp. 4433–4478, 2023.
- [3] S. G. Rizzo, M. Brucato, and D. Montesi, "Ranking Models for the Temporal Dimension of Text," *ACM Trans. Inf. Syst.*, vol. 41, no. 2, pp. 1–34, 2023.
- [4] E. Zangerle and C. Bauer, "Evaluating Recommender Systems: Survey and Framework," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–38, 2023.
- [5] R. Sivasankari and J. Dhilipan, "Hybrid scientific article recommendation system with COOT optimization," *Data Sci. Manag.*, vol. 7, no. 2, pp. 99–107, 2024.
- [6] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intell. Data Anal.*, vol. 21, no. 6, pp. 1487–1524, 2017.
- [7] A. Arsad, M. Hamid, and M. Santosa, "PENERAPAN TEKS MINING DAN COSINE SIMILARITY UNTUK MENENTUKAN KESAMAAN DOKUMEN SKRIPSI," *IJIS - Indones. J. Inf. Syst.*, vol. 9, no. 1, 2024.
- [8] M. Amin, A. Firdaus, and Y. Dwitayanti, "Model Rekomendasi Jurnal dengan Algoritma Jaccard Similarity dan Protokol OAI-PMH," *J. Pendidik. Dan Teknol. Indones.*, vol. 4, no. 10, pp. 489–499, 2025.
- [9] D. El Alaoui, J. Riffi, M. A. Sabri, B. Aghoutane, A. Yahyaouy, and H. Tairi, "Comparative Study

- of Filtering Methods for Scientific Research Article Recommendations,” *Big Data Cogn. Comput.*, vol. 8, no. 12, 2024.
- [10] İ. KUŞ, S. B. Keser, and S. Okay, “A Novel Article Recommendation System Empowered by the Hybrid Combinations of Content-Based State-of-the-Art Methods,” *Int. J. Appl. Math. Electron. Comput.*, vol. 11, no. 1, pp. 1–12, 2023.
 - [11] A. P. Putra, D. P. S. Putri, and A. A. K. A. C. Wiranatha, “Scientific Paper Recommendation System: Application of Sentence Transformers and Cosine Similarity Using arXiv Data,” *J. Appl. Informatics Comput.*, vol. 9, no. 4, 2025.
 - [12] U. Zachrias and W. Gunawan, “Classification of Document Similarity Using Jaccard Coefficient Approach Winnowing Algorithm with Indikasi Plagiarisme,” *Sci. Inf. Technol. Appl.*, vol. 1, no. 1, 2025.
 - [13] A. Y. Bramantya, T. Hasanuddin, and F. Umar, “Analisis Metode Winnowing Dalam Pendeteksian Plagiarisme Judul,” *Bul. Sist. Inf. dan Teknol. Islam*, vol. 3, no. 4, 2022.
 - [14] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, p. 6086, 2024.
 - [15] R. Klemiński, P. Kazienko, and T. Kajdanowicz, “Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification,” *J. Inf. Sci.*, vol. 48, no. 3, pp. 349–373, 2022.
 - [16] K. W. Boyack and R. Klavans, “Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, 2010.
 - [17] P. Sjögarde and P. Ahlgren, “Seed-based information retrieval in networks of research publications: Evaluation of direct citations, bibliographic coupling, co-citations, and PubMed-related article score,” *J. Assoc. Inf. Sci. Technol.*, vol. 75, no. 13, 2024.
 - [18] E. M. Lghaouch, F. Z. Abbour, S. Ounacer, S. Ardchir, and M. Azzouazi, “Enhancing Recommendation Systems via Aspect-Level Sentiment Analysis for Identifying Unreliable Reviews,” *Indones. J. Electr. Eng. Informatics*, vol. 14, no. 2, 2026.
 - [19] E. Yuniar, D. S. Utsalinah, and D. Wahyuningsih, “Implementasi Scrapping Data Untuk Sentiment Analysis Pengguna Dompok Digital dengan Menggunakan Algoritma Machine Learning,” *J. Janitra Inform. dan Sist. Inf.*, vol. 2, no. 1, 2022.
 - [20] A. Rahmatulloh and R. Gunawan, “Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar,” *Indones. J. Inf. Syst.*, vol. 2, no. 2, 2020.
 - [21] T. M. Fahrudin, P. A. Riyantoko, and K. M. Hindrayani, “Implementation of Web Scraping on Google Search Engine for Text Collection Into Structured 2D List,” 2023.
 - [22] F. P. Utama, T. Mardiansyah, R. Faurina, and A. Vatesia, “SCIENTIFIC ARTICLES RECOMMENDATION SYSTEM BASED ON USER’S RELATEDNESS USING ITEM-BASED COLLABORATIVE FILTERING METHOD,” *J. Tek. Inform.*, 2023.
 - [23] A. Zamsuri, S. Defit, and G. W. Nurcahyo, “Classification of Multiple Emotions in Indonesian Text Using The K-Nearest Neighbor Method,” *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 2, 2023.
 - [24] C. Cheliger, J. Huang, G. Wu, N. Bhuiyan, Y. Xu, and Y. Zeng, “Machine learning in requirements elicitation: a literature review,” *Artif. Intell. Eng. Des. Anal. Manuf.*, 2022.
 - [25] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, “Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation,” *J. Big Data*, vol. 8, no. 1, 2021.
 - [26] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, “Flexibility of Indonesian text pre-processing library,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 1, 2019.
 - [27] I. O. Suzanti and A. Jauhari, “Comparison of Stemming and Similarity Algorithms in Indonesian Translated Al-Qur’an Text Search,” *J. Ilm. Kursor*, vol. 11, no. 2, p. 91, 2022.
 - [28] M. Johari and A. D. Laksito, “The Hybrid Recommender System of the Indonesian Online Market Products using IMDb weight rating and TF-IDF,” *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 5, no. 5, 2021.
 - [29] W. Yulita, M. C. Untoro, M. Praseptiawan, I. F. Ashari, A. Afriansyah, and A. N. C. Pee, “Automatic Scoring Using Term Frequency Inverse Document Frequency Document Frequency and Cosine Similarity,” *Sci. J. Informatics*, vol. 10, no. 2, 2023.
 - [30] A. Widiyanto, E. Pebriyanto, F. Fitriyanti, and M. Marna, “Document Similarity Using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity,” *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 4, no. 2, 2024.
 - [31] S. Schleimer, D. S. Wilkerson, and A. Aiken, “Winnowing,” 2003.

- [32] D. Nasien and E. P. Yansen, "Implementation of the Wining Algorithm in Detecting Plagiarism in Title and Abstract of Student's Final Project," *IT J. Res. Dev.*, vol. 7, no. 1, pp. 23–35, 2022.
- [33] T. Wahyuningsih, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *J. Appl. Data Sci.*, vol. 2, no. 2, 2021.
- [34] S. Purwaningrum, A. Susanto, and N. W. A. Prasetya, "Comparison of Dice Similarity and Jaccard Coefficient Against Wining Algorithm For Similarity Detection of Indonesian Text Documents," *J. Appl. Intell. Syst.*, vol. 6, no. 1, pp. 10–22, 2021.
- [35] N. Haubner and T. Setzer, "Applying Optimal Weight Combination in Hybrid Recommender Systems," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.
- [36] H. Arabi and M. Akbari, "Improving plagiarism detection in text document using hybrid weighted similarity," *Expert Syst. Appl.*, vol. 207, p. 118034, 2022.
- [37] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph Neural Networks in Recommender Systems: A Survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, 2023.
- [38] R. Widayanti, "Improving Recommender Systems using Hybrid Techniques of Collaborative Filtering and Content-Based Filtering," *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 289–302, 2023.
- [39] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, vol. 119, no. 546, pp. 1434–1445, 2024.
- [40] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, Dense, and Attentional Representations for Text Retrieval," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 329–345, 2021.
- [41] D. Yang, D. Zhu, H. Gai, and F. Wan, "Semantic Similarity Calculating based on BERT," *J. Electr. Syst.*, vol. 20, no. 2, pp. 73–79, 2024.
- [42] A. D. P. Ariyanto, F. K. Fikriah, and A. F. Setyawan, "Impact of Statistical and Semantic Features Extraction for Emotion Detection on Indonesian Short Text Sentences," *CommIT (Communication Inf. Technol. J.)*, vol. 19, no. 1, pp. 1–13, 2025.